HDAT9700

Missing data and multiple imputation

Online tutorial

UNSW SYDNEY Australia's Global University

UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

Outline today

1. The problems caused by missing data
2. Tips for exploring missing data
3. Types of missing data
4. Single imputation methods
5. Multiple imputation
6. MI via iterative chained equations

UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

The problem of missing data

1. **Reduced N (equals wider variance)**
   - Can be ameliorated by collecting more data
   - Often less of a concern in big data era (e.g. when using matching we happily throw away data to get better estimates)

2. **Potential for bias**
   - More concerning
   - Arises **if** individuals with missing data are **different** from those with complete data
   - Can be corrected, but needs observed data + assumptions

Activity

- Missing data is not much of a problem if:
  1. The proportion of missing data is small
  2. There is no difference between individuals with missing data and complete data

- See illustration in tutorial

Suggestions for exploring missing data

Before diving into complex methods like multiple imputation it is vital to have a good understanding of your data

What is your approach? Discuss!

1. Refer to existing documentation and **talk** to data custodians/users
2. Look at the proportion missing for each variable (**is.na()** function)
3. Look at proportion with complete cases versus 1 or more incomplete variables (**mice::cci()** function)
4. Check for trends or **patterns** in missing data (time/geography/key vars)
5. Compare complete background variables for complete versus incomplete cases

UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

Missing data mechanisms

A taxonomy of three types of missing data

1. Missing Completely at Random (MCAR)
2. Missing not at random (MNAR)
3. Missing at Random (MAR)

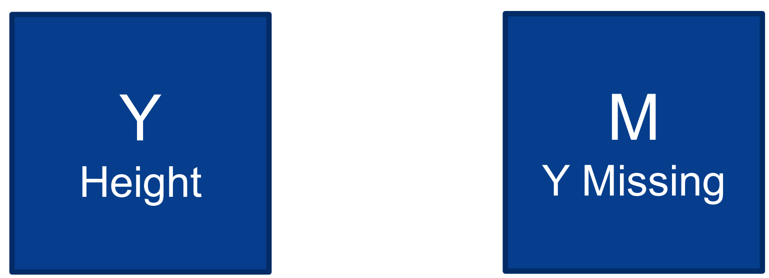
Prof. Donald Rubin

(also proposed propensity score matching)

Image from harvard.edu

Missing Completely at Random (MCAR)

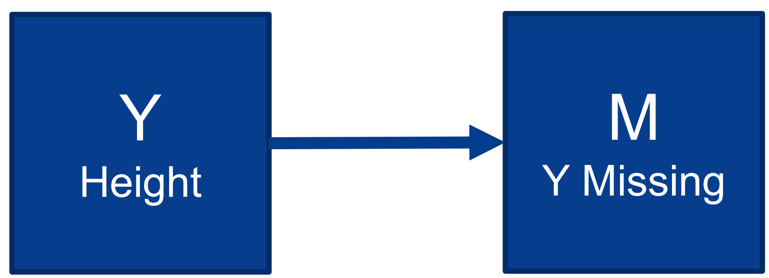
- The probability of being missing is independent of the data value
- The missingness is truly “random”
- e.g. a random sample of a Wave 1 cohort members were chosen for follow-up at wave 2
- Complete case analysis has higher variance but unbiased!



Missing Not at Random (MNAR)

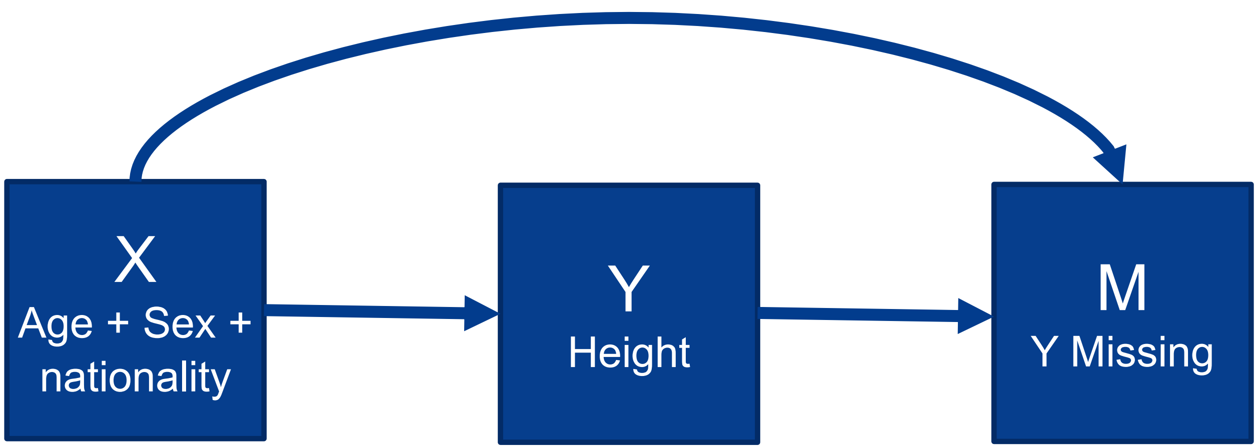
- The probability of being missing does depend on the data
- Missing cases are different to complete cases

- e.g. Sicker participants might be more likely to drop out of an RCT
- Complete case analysis will have higher variance and biased!



Missing at Random (MAR)

- The probability of being missing is independent of the data value, **conditional on the observed data**
- Missingness is random within categories defined by the observed variables
- Complete case analysis is unbiased if you control for X
- **Most approaches to address missing data assume a MAR scenario**



Review

Can you think of a practical example of each of these three scenarios?

1. Missing completely at random (MCAR)

Being observed or not is truly random in the conversational sense of random

2. Missing not at random (MNAR)

*Being observed or not depends on the **unobserved** data*

3. Missing at random (MAR)

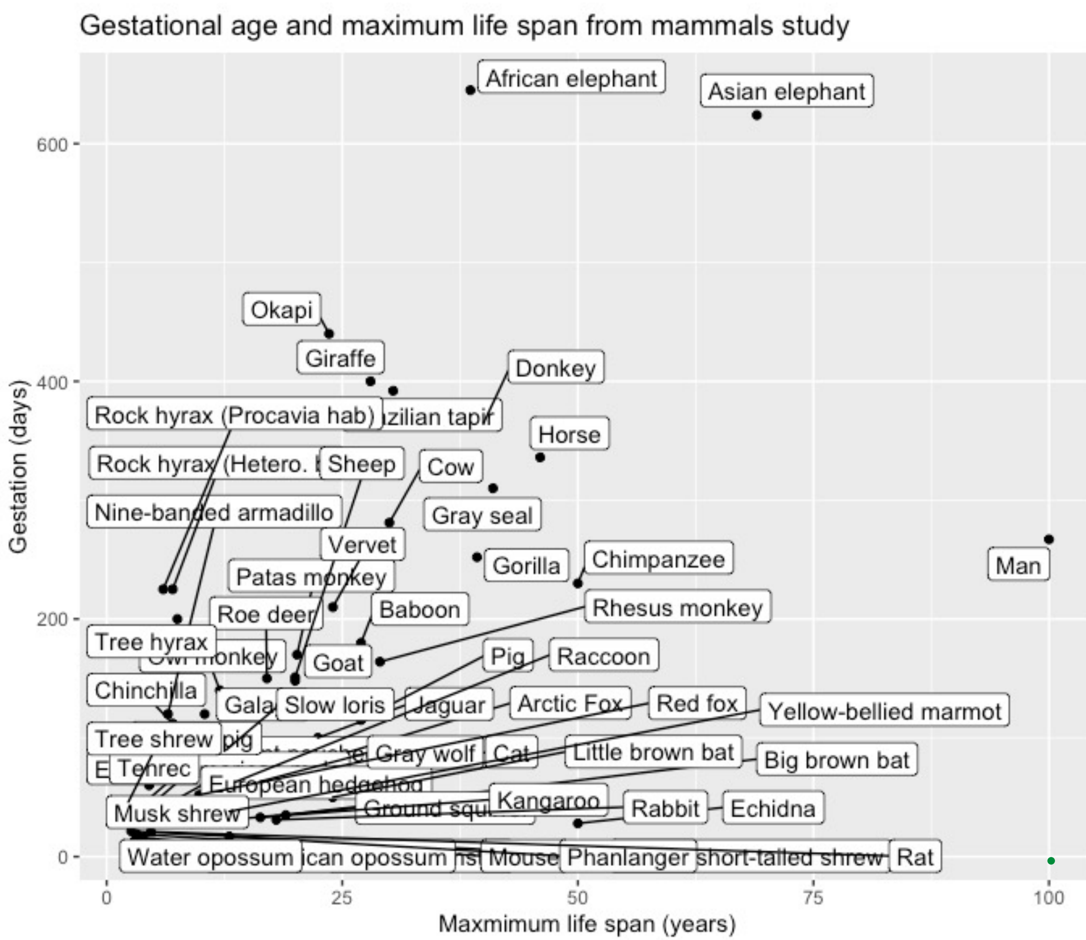
Being observed or not depends on observed data.

Single imputation

Two models: imputation model and analysis model

- Mean
- Regression
- Stochastic regression
- Predictive mean matching
- Last observation carried forward

Single imputation methods



Illustrated using mammal sleep data from VIM (*VIM::sleep*) and mice packages (*mice::mammalsleep*)

UNSW SYDNEY

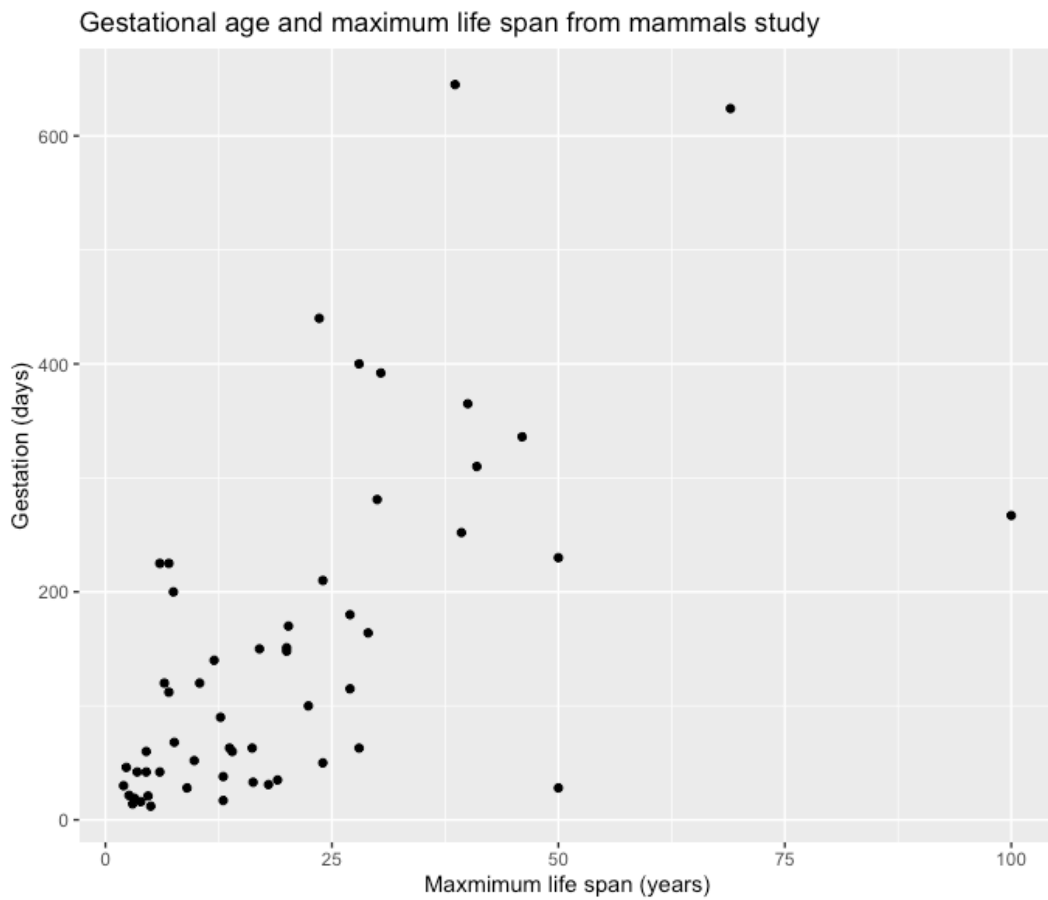
CENTRE FOR BIG DATA RESEARCH IN HEALTH

Single imputation methods

Illustrated using mammal sleep data from VIM (*VIM::sleep*) and mice packages (*mice::mammalsleep*)

Single imputation methods

Mean imputation



UNSW SYDNEY

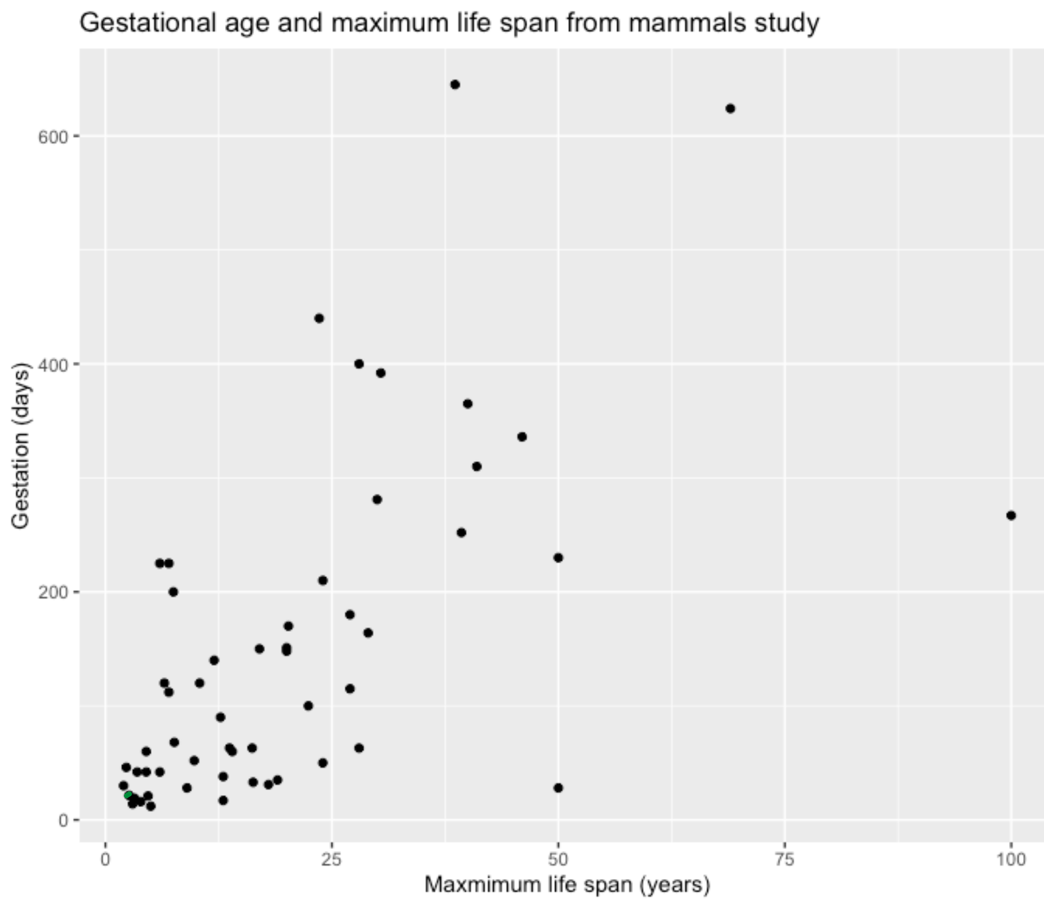
CENTRE FOR BIG DATA RESEARCH IN HEALTH

Single imputation methods

Regression imputation

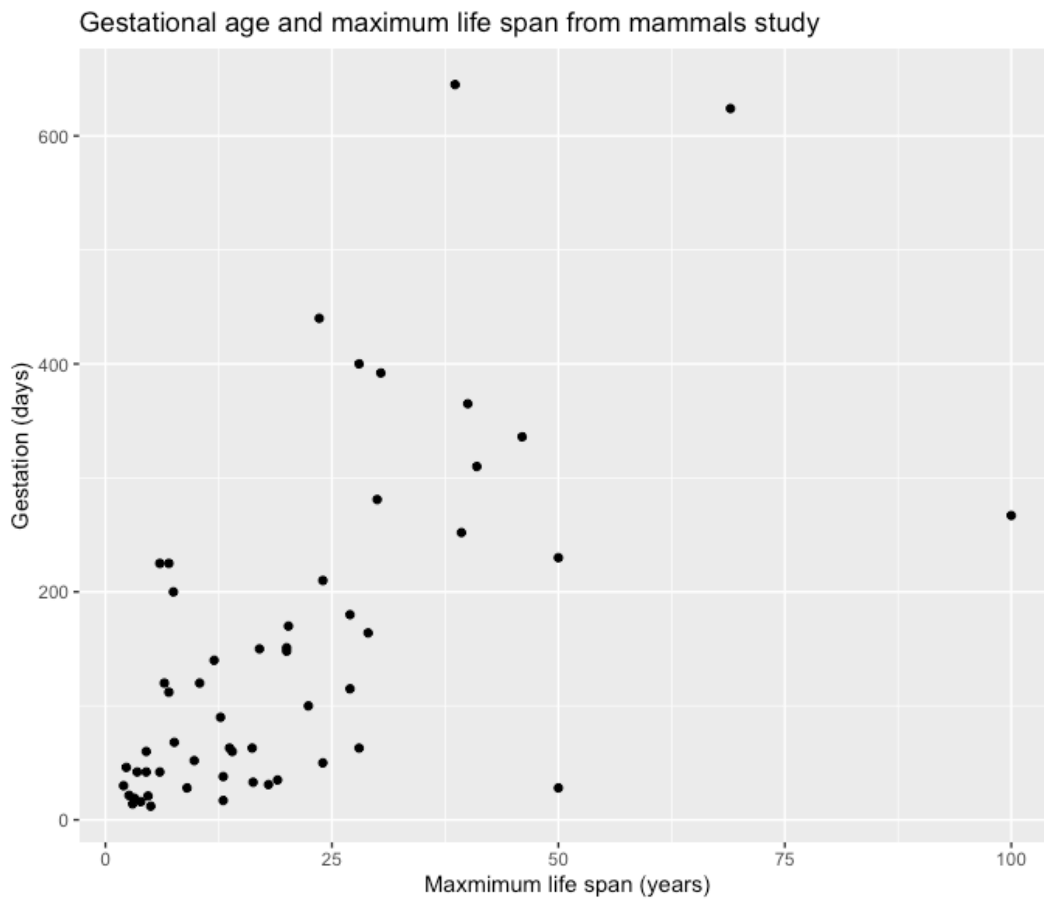
Single imputation methods

Stochastic regression imputation



Single imputation methods

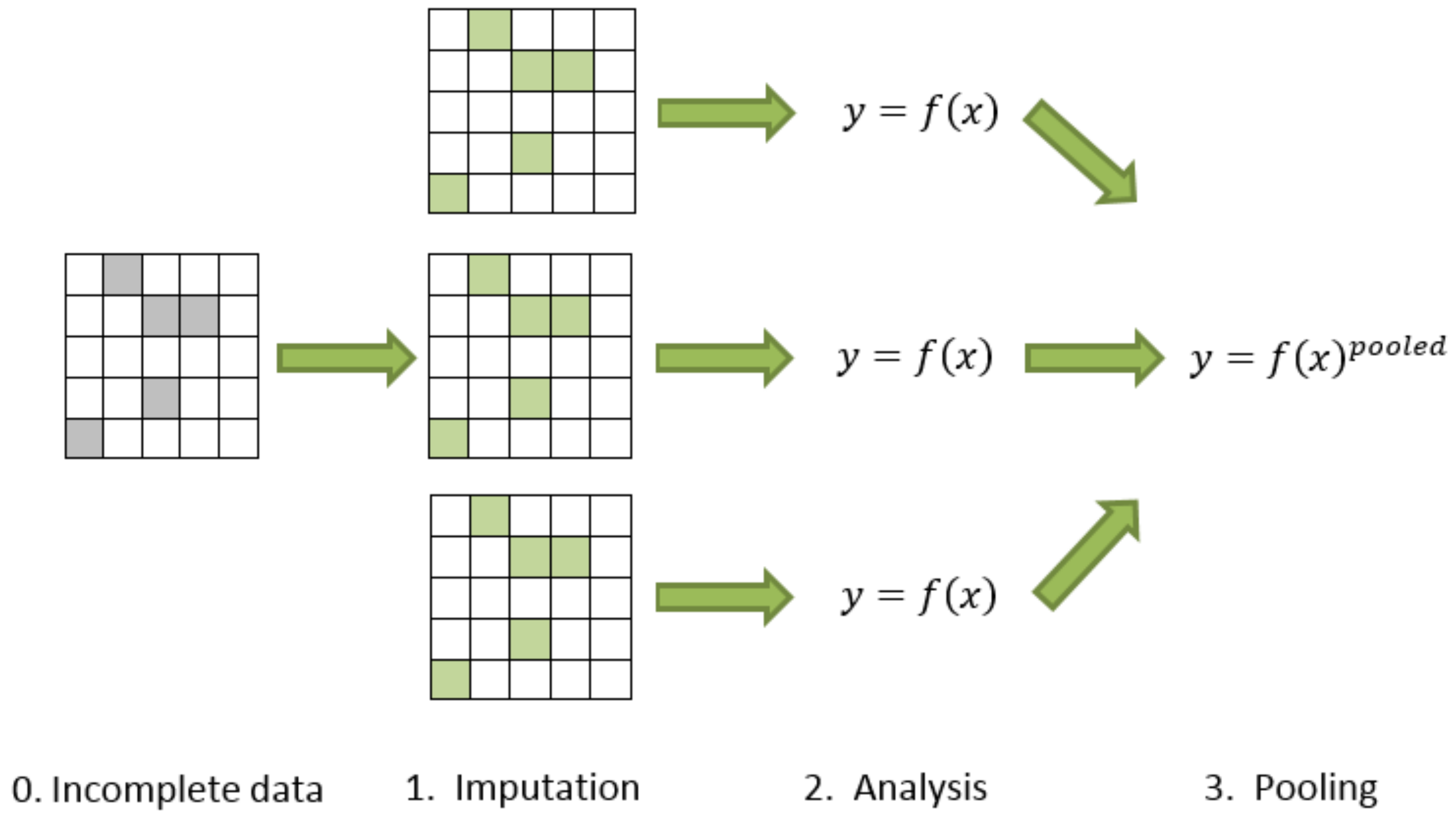
Predictive Mean Matching imputation



Visit Moodle to complete the myExperience survey or scan the QR code to complete on your phone

Tell us about your experience.
Shape the future of education at UNSW.

Multiple Imputation is a three-stage process



Rubin's Rules for combining Estimates base on multiple imputation

Point estimate of a parameter of interest: denoted Q

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_l$$

The average of the m estimates – easy!

Rubin's Rules for combining Estimates based on multiple imputation

Variance of the parameter of interest: T

$$T = \bar{U} + B + \frac{B}{m}$$

1. Within Variance

$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m} \hat{U}_l$$

2. Between Variance

$$B = \frac{1}{m-1}\sum_{l=1}^{m} \left(\hat{Q}_l - \bar{Q}\right)^2$$

3. Simulation Variance

$$\frac{B}{m}$$

Average variance across m imputations

Variation of variance across m imputations

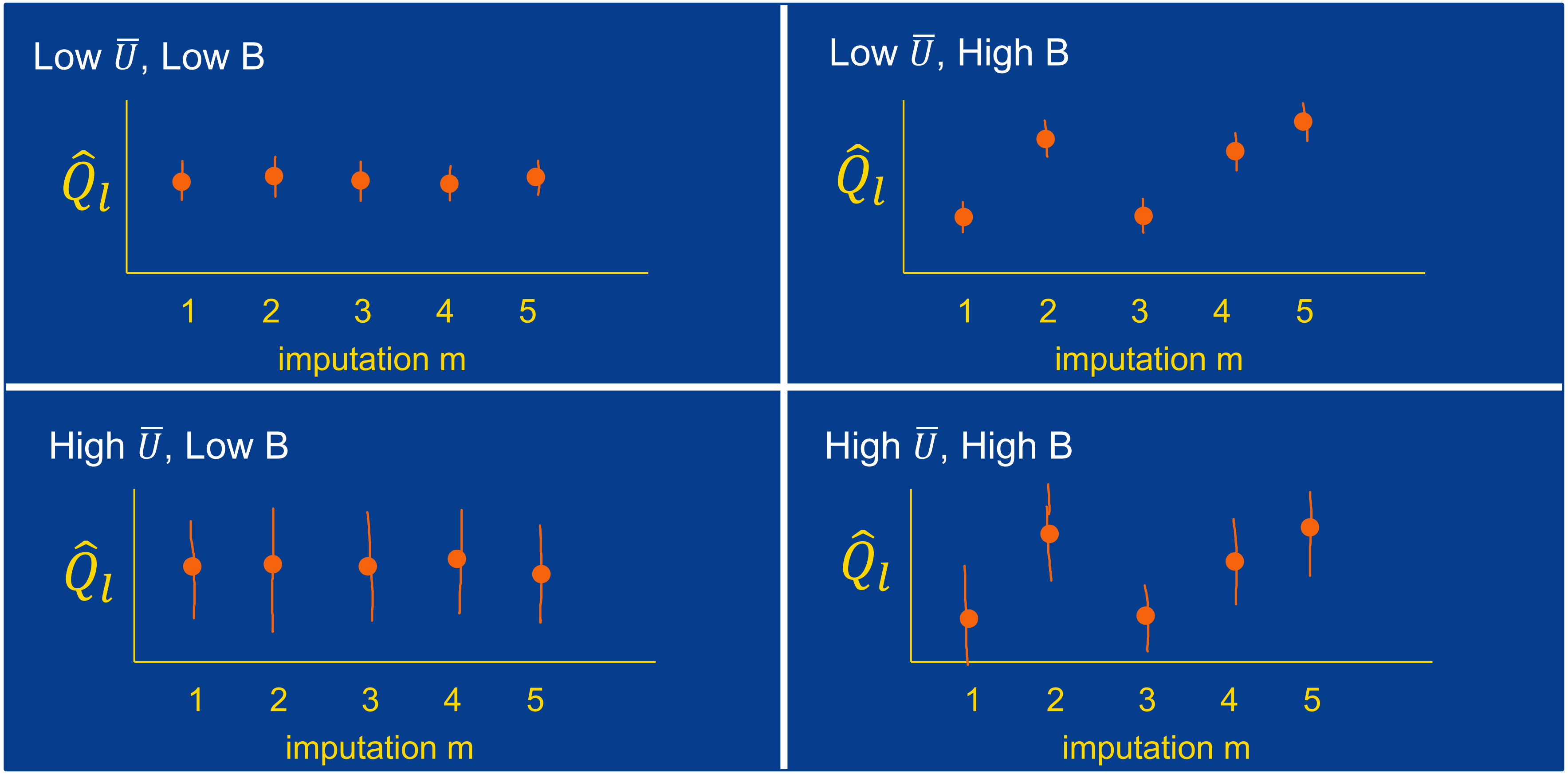
Accounts for finite m

UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

Rubin's Rules for combining Estimates base on multiple imputation

Variance of the parameter of interest: T



Iterative Chained Equations (ICE)

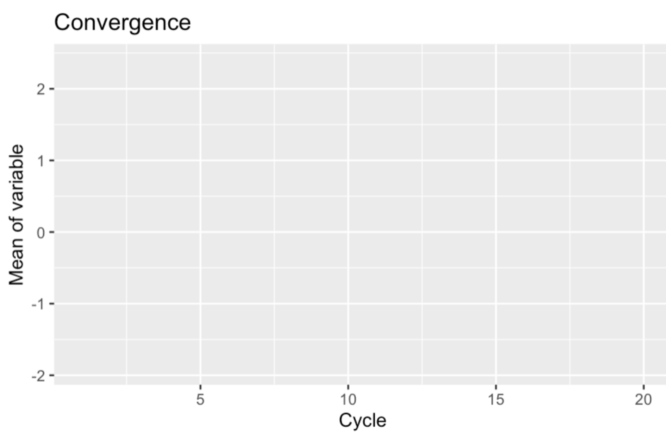
- **Why ICE?**
  - Need to draw from joint distribution
  - Difficult when non-normal data or many different data types (binary, categorical, count etc)
  - ICE achieves draws from correct joint distribution by **iteratively drawing from univariate distributions**

Iterative Chained Equations (ICE)

- **How does it work?**
  - Start by filling in all incomplete variables with some plausible value
  - Fit a model for each incomplete variable in turn, using all remaining variables as predictors.

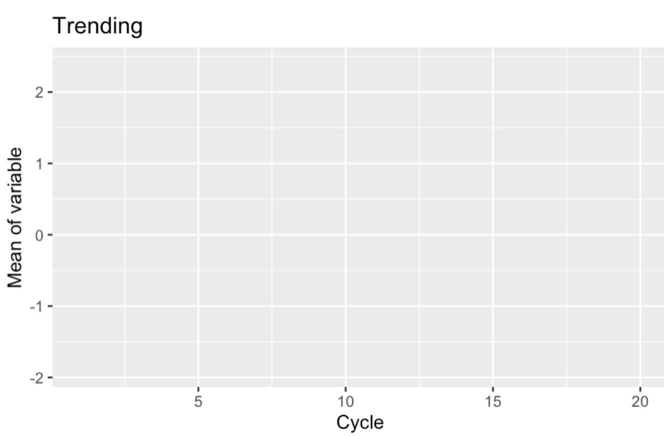
Iterative Chained Equations (ICE)

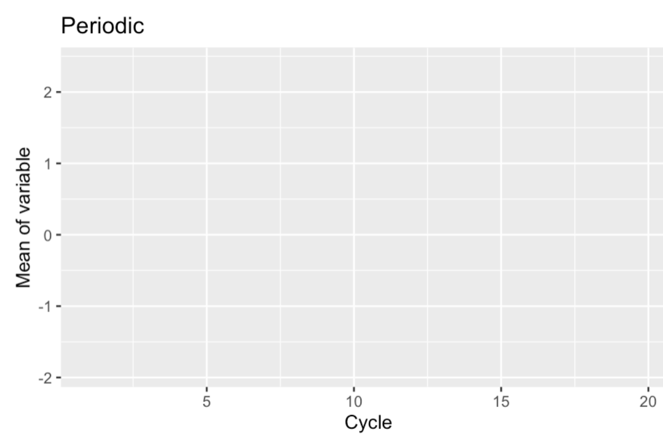
- **How many cycles?**
  - Usually 10-20 are adequate
  - Examine trace plot for convergence



On convergence:

- No trends
- Free mixing across parameter space





CENTRE FOR
BIG DATA RESEARCH
IN HEALTH

Iterative Chained Equations (ICE)

- **Features and advantages**
  - Valid imputations from intractable multivariate distributions
  - Specify the appropriate model type for each variable
  - Tailor model predictors (*minimise noise & overfitting, e.g. time series data*)
  - Passive imputation (*maintain fixed relationships e.g. BMI*)
  - Conditional imputation (*e.g. # cigarettes per day only for smokers*)

Reporting the results of multiple imputation

For an example of how to report the methods and results for a regression analysis please see Supplement A Missing Data and Multiple Imputation in

Hanly M, Falster K, Banks E, et al. **Role of maternal age at birth in child development among Indigenous and non-Indigenous Australian children in their first school year: a population-based cohort study.** *Lancet Child Adolesc Health* 2019 http://dx.doi.org/10.1016/S2352-4642(19)30334-7.

Link to supplement here: https://ars.els-cdn.com/content/image/1-s2.0-S2352464219303347-mmc1.pdf

Summary

Missing data can increase **variance** and bias in model estimates

The magnitude of bias depends on the % missing data and the association between the parameter of interest and the probability of being missing

Need to **understand mechanisms** leading to missing data

- **Talk** to data custodians!
- **Summarise** patterns of missing data

Single imputation ok for small % missing data

Multiple imputation

- Impute multiple times (Usually 5-20)
- Estimate model in each complete dataset
- Average model estimates